Server Consolidation Techniques in Virtualized Data Centers: A Survey

A. Varasteh, and M. Goudarzi, *Member, IEEE*
Computer Engineering Department, Sharif University of Technology, Tehran, I.R.Iran

***Abstract*— Data Centers and their applications are growing exponentially. Consequently, their energy consumption and environmental impacts have also become increasingly more important. Virtualization technologies are widely used in modern data centers to ease the management of the data center and also reduce its energy consumption. Data centers that employ virtualization technologies are typically called *virtualized or cloud data centers*. Virtualization technologies enable Virtual Machine (VM) Live Migration which allows the VMs to be freely moved among Physical Machines (PMs) with negligible downtime. Thus, several VMs can be packed on a single PM so as to let the PM run in its more energy efficient working condition. This technique is called *Server Consolidation* and is an effective and widely used approach to reduce total energy consumption in data centers. Server consolidation can be done in various ways and by considering various parameters and effects. This paper presents a survey and taxonomy for server consolidation techniques in cloud data centers. Special attention has been devoted to the parameters and algorithmic approaches used to consolidate VMs onto PMs. In this end, we also discuss open challenges and suggest areas for further research.**

***Index Terms*—Cloud computing, data center, energy efficiency, server consolidation, virtualization, resource allocation.**

I. Introduction

Cloud computing and data centers have become an important part of our daily lives because of various internet-scale services, such as internet-wide search and email services, that we have got used to take advantage of on a regular basis [1]. Cloud datacenters can provide the illusion of unlimited resources to the users through the Internet, and big companies such as Amazon [2], Microsoft [3], Google [4] and IBM [5] are developing and providing cloud-based services for their customers. The United States Environmental Protection Agency (EPA) has reported that the energy used by federated servers and data centers was about 100 billion KWh in 2011 [6]. Also, world electricity demand for data centers is expected to increase by more than 66% over the period 2011-2035 [7]. This high energy consumption of data centers along with their high rate of growth and total energy and carbon footprint has made it inevitable to apply Green computing techniques and reduce data center energy consumptions for sustainable growth. Virtualization technology provides several features and benefits to cloud providers such as resource multiplexing, live migration, server consolidation and VM resizing [8]. Using these features, cloud providers can provide mostly unlimited and on-demand resources to their customers.

Due to unpredictable and growing demand for resources, data centers need to offer high performance computation and large volume data storages [9]. These physical resources along with air conditioning and cooling equipment are the main power consumers in data centers. Moreover, the resource utilization is one of the most important factors that affects data center energy consumption [10]. Measurements have shown that in data centers, average server utilization is between 10% and 50% [11]. This can waste lots of energy because typically an idle server consumes as large as 50% of a fully utilized server [12]. Consequently, one way to reduce the data center energy waste is server consolidation technique where data center VMs are packed on fewer number of Physical Machines (PM). This technique is based on virtualization technology and by using it, one can increase the server utilization, and moreover, can put now free PMs into standby mode to more effectively reduce the energy consumption of data center [13].

Server consolidation techniques pack a number of VMs on fewer number of PMs to optimize the resource utilization and reduce the power consumption by letting the PMs run in optimally efficient energy and more energy proportional state. The important feature that makes the server consolidation technique even more attractive is VM Live Migration. Using VM live migration one can transfer a running VM from a PM to another PM without considerable service downtime. This consolidation can be done in different ways considering various parameters.

We believe that it's hardly possible to produce a completely accurate survey and classification of server consolidation techniques that is doing justice to every viewpoints. In this paper we present a survey on different consolidation techniques, the different parameters their considered, their objectives they pursued on the datacenter operations and costs (e.g. energy consumption, cooling efficiency), and the different optimization methods they used to solve the consolidation problem. Also we presented some open challenges and areas for further research. We discuss the server consolidation problem from several aspects that are mostly affect the problem and the way that the problem is solved and evaluated. In the first criteria, we discuss the time of decision making: static, dynamic and dynamic with load prediction. Static techniques is used where there is a stable set of VMs and predictable demand patterns as opposed to dynamic consolidation techniques which are used in data centers with time-varying and stochastic

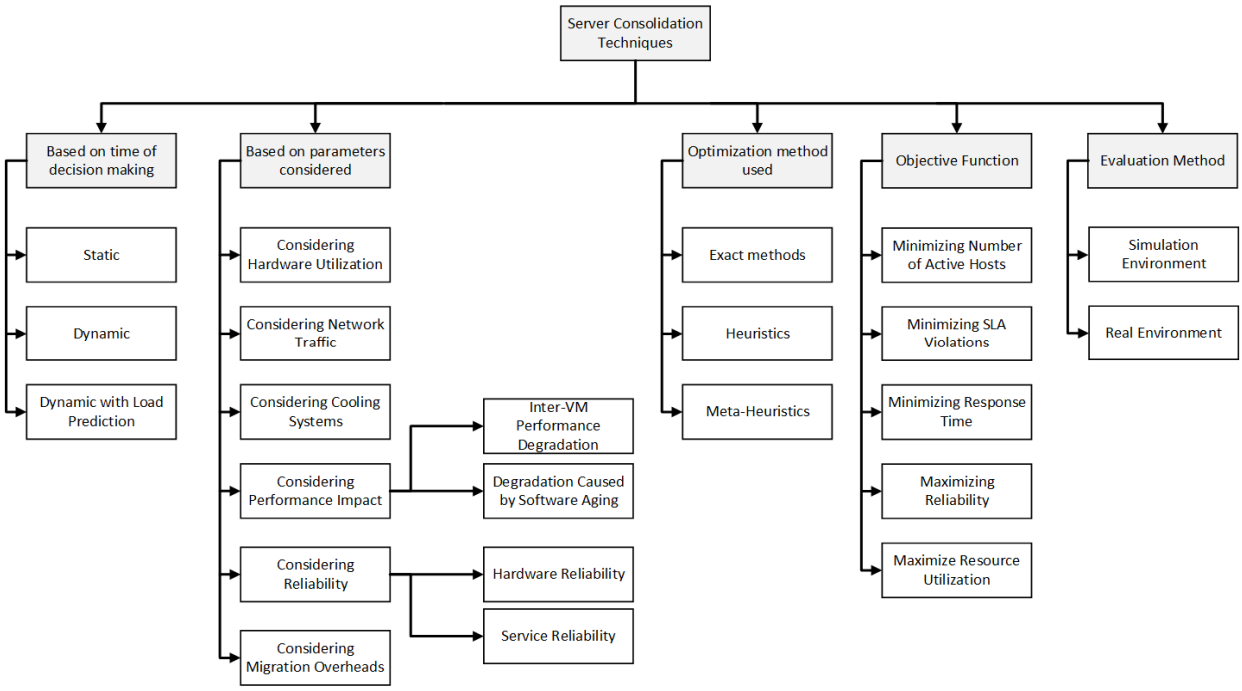

Fig. 1. An overview of our classification of consolidation techniques.

workloads [14]. Also, another method is dynamic techniques with load prediction which can be quite successful in certain predictable applications. Another aspect is parameters which are considered in server consolidation problem. These parameters are affect the data center operational costs and user experience etc. As we know, there are several parameters that could be considered into the account: : hardware utilization, data center network traffic, data center cooling systems, performance of running applications, reliability of the data center devices and also migration overheads. Each of these parameters can highly affect the server consolidation process.

Two other aspects we used for classification are the optimization method and the objective function employed in the optimization process. Above parameters are used to effectively achieve objectives such as minimization of energy, overheads (both network and computational) and costs, and maximization of application performance, hardware reliability and service reliability/availability. These various objective functions are presented in table 3 with other considered parameters in the optimization problem. Our fourth discussed aspect is optimization methods used in order to solve the server consolidation problem. Since the server consolidation problem can be mapped to a high-dimensional NP-Hard bin packing problem, it is often formulated and solved using various heuristics and meta-heuristics. Finally, another important aspect to classify the approaches is the way they are evaluated: in simulation environment or in real data centers. Approaches that have been implemented and evaluated in real data centers are more reliable in terms of their claimed achievements than simulated approaches. Under simulated environments, the reported evaluations need to carefully consider limitations, assumptions, and precision of the employed simulation environment. These evaluation methods are also given in table 3 which presents an overall view and comparison of prior works. Fig. 1 gives an overview of the presented classification followed and detailed in the rest of this paper.

The remainder of his paper is organized as follows: Section 2 presents the system model for resource allocation in data centers. In section 3 we present the classification based on time of decision making of the server consolidation algorithm. In section 4, techniques are classified based on the parameters taken into account by the consolidation approach, and in section 5, a classification based on the employed optimization method is presented, followed by conclusion and open challenges in section 6.

II. System Model

Typical architecture of a data center resource allocation system is presented in Fig. 2. Each PM runs a *Virtual Machine Monitor* (VMM) e.g. Xen [15], and one or more virtual machines. Each virtual machine runs an application or an application component. Each PM communicates with the data center manager system. Data center manager comprises several components: *controller*, *monitoring engine*, *migration manager*. Monitoring engine continuously gathers processor, network interface, memory usage and other data for each PM through the controller. It then processes the data and passes the data and statistics to the migration manager component. Migration manager then uses the information and the consolidation algorithm to determine the migrations that have to be done. It then applies the changes to the data center configuration using the controller. There is also another component that some of data center management systems use: the *predictor* which predicts the future workload to help the

migration manager generate a better configuration for data center.

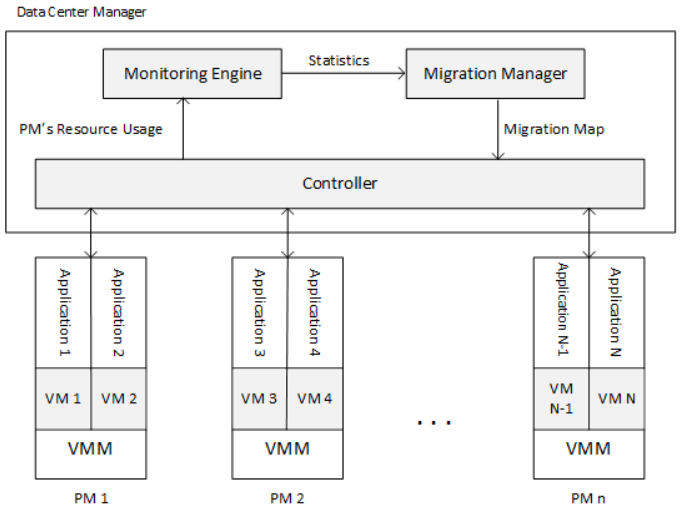

Fig. 2. A typical system model for a data center resource allocation system

III. Classification Based on Time of Decision Making: Static vs. Dynamic

Server consolidation maps the VMs to proper PMs considering several parameters and limitations. Several virtualization vendors provide management tools [16] with some third-party tools [17] for consolidation and data center management. Server consolidation can be done in two ways: Static and Dynamic which will be discussed in below.

A. Static Consolidation Techniques

In static server consolidation, VM to PM mappings are not changed for a long time, and no migration is done with workload changes during that time [18]. An advantage of this approach is in batch job processing and applications with consistent demands. A disadvantage of static consolidation is resource overprovisioning. In static consolidation, the resources are allocated in a way that satisfies the peak load demands, and consequently, most of the time they are wasted most of the time when the VM is not working at its peak load similar to the traditional data centers case. Kishaly Halder *et al.* in [19] introduce an algorithm using static consolidation which tries to generate the initial placement and the resources amount for VMs considering energy consumption. Speitkamp *et al.* in [20] used a multidimensional bin-packing formulation (MBP) to model the problem of static and dynamic server consolidation. Wolke *et al.* in [14] argued the static consolidation method against the dynamic resource allocation techniques which are compared in a real data center experiment.

B. Dynamic Consolidation Techniques

In dynamic consolidation, the consolidation algorithm is run in response to workload variations or at specific time intervals and may decide to migrate VMs to other PMs [21]. Most of studies are done on the second method, dynamic consolidation. As we mentioned earlier, dynamic consolidation algorithm will run in specific time periods or specific events choice of which can affect the algorithm efficiency. The algorithm presented by Gergo Lovasz *et al.* in [22] runs every 10 minutes. Increasing or decreasing the time period could affect the resource utilization, energy consumption and/or data center performance. Running the algorithm in short periods makes the data center changes happen more rapidly, and hence the servers will turn on and off more than before and as a result, the servers' life time will decrease. Also, more bandwidth has to be allocated to VM migrations and the real users' network traffic will face lower available bandwidth. On the other hand, longer time intervals also have negative impacts on data center performance e.g., due to not reacting quickly enough to workload changes, servers might be overloaded and this can reduce the application performance and may violate the SLA. Also, too long consolidation periods may lose energy saving opportunity, because during the time period, several servers could be potentially ready to go to standby mode, but we find them too late and until then, they stay running and consume energy [23], [24], [25].

In addition to time periods, there is another way to determine when to run the algorithm i.e. at specific events. Wei Deng *et al.* in [26] define a trigger to run the algorithm. The definition is based on a load parameter which consists of some multidimensional resource (i.e. CPU, Memory, Disk and Network) utilizations. When load parameter reaches to a specific value, the consolidation algorithm runs and prevent possible server(s) from over- or under-loading. Also, John J. Prevost *et al.* in [23] presented a stochastic optimization model which determines the optimal update frequency for changing the VM to PM mapping.

C. Dynamic Consolidation with Load Prediction Techniques

One of the main reasons that energy consumption of data centers are very high is because servers are online but are idle. To save power, the servers must be switched to lower power states when they are not in use. Also, switching a server from a power state to another causes delay and energy overheads. So, if a server will not be needed for a long time, it worth to keep the server off rather than turn it on and cause unnecessary energy and time overheads. These facts rise the need for prediction techniques which can be used to estimate the future data center workload. We can use these prediction techniques to properly decide when and for how long a server need to be turned off or be awake to process new VM requests. Several works have used these techniques to provide efficient and desirable server consolidation algorithms. Wei Xu *et al.* in [27] presented three different prediction algorithms: standard autoregressive (AR) model which account for temporal correlations between the current value of a parameter and its history, a combined ANOVA-AR model that combines AR method and analyzed long-term repeatable patterns in a time series, as well as a multi-pulse (MP) model. MP was first used in speech processing which analyze both long-term and short-term pattern in an online manner. Zhenhuan Gong *et al.* in [28] proposed *Predictive Elastic Resource Scaling (PRESS)*

framework. *PRESS* tries to allocate just enough resources to VMs in a way that minimize SLA violations and resource wastages. It tracks the dynamic VM demands and predicts these resource requirements in the near future using light-weight signal processing and statistical learning techniques. However, predicting data center workloads can be very complex and challenging due to the diversity and stochastic arrivals of client requests, while each coming at a different time and requesting different amounts of resources (CPU, memory, bandwidth, etc.).

IV. CLASSIFICATION BASED ON THE CONSIDERED PARAMETERS

A. Considering Hardware Utilization

One of the most used parameters in server consolidation algorithms and resource provisioning is hardware utilization. Various hardware resources (e.g., CPU, memory, disk, and network) could be considered in the optimization algorithm. Several studies [29], [13], [30] consider only CPU for the proposed algorithm, while in other studies the number of considered resources is increased which potentially leads to better mappings. Authors in [31], [32] considered CPU and Memory as optimization parameters. Also, Beloglazov *et al.* in [33] considered CPU, memory and network utilization as optimization parameters and in [26], authors consider CPU, memory, network and disk. In the mentioned studies, authors used the resource utilizations to model and solve the optimization problem. Fox *et al.* in [34], used 13 different metrics to model the VM and server performance and based on it, the resource provisioning is done. Also, the objective function could be different work by work, but maximizing these resource utilizations is one of the important objective functions in many works [35], [36]. But the goal is not simply 100% of utilization, and the server should have buffers due to tolerating the fluctuation of workloads.

B. Considering Network Traffic

Many studies on server consolidation use the resource (e.g., CPU, Memory, I/O) utilization to model the problem, but neglect an important factor, namely network traffic and the communication between VMs, that can change the best PM to host a VM. Traffic flows on the data center network and also the relations between VMs and communication among them could affect the data center performance and Quality of Service (QoS). Also in batch processing jobs, delay and long communications time between two nodes can affect the task completion time and hence increases the data center energy consumption. Also, this can lead to situations that VM pairs and heavy traffic between them are placed on PMs which are far from each other (e.g., different racks) and incur large network traffic cost between them [37]. Xiaoqiao Meng *et al.* in [37] proposed an algorithm that considers data center network topology and network traffic patterns to increase data center service performance. This algorithm takes traffic matrix among VMs and communication cost matrix among PMs, as inputs and produces VM to PM mapping such that the traffic passing through the switches is minimized as the output.

Network traffic in production data centers is very bursty and fluctuating [37], [38]. So, it would not be easy to have a reliable and deterministic estimation for bandwidth demand. Meng Wang *et al.* in [39] used random variables to characterize the future bandwidth usage and modeled the problem using stochastic bin packing and considered bandwidth demand for each VM. Also, they considered network capacity of a PM as a limitation for the optimization problem. Min Cut Ratio-aware VM Placement (MCRVMP) is proposed in [40] which in addition to resource demands such as CPU and memory, the VM communication demands is also considered in the proposed optimization problem. Using all this, the proposed algorithm tries to minimize the maximum ratio of the demand and the capacity across all cuts in data center network. Also, it handles the unpredicted traffic bursts by allocating some spare capacity on each network cut.

Data intensive jobs in data centers need lots of storages and computation resources and produce heavy data streams [41]. Kliazovich *et al.* in [42] proposed an algorithm with two main functions showing a trade-off between them which should be optimized. These two functions are: 1) Server consolidation to minimize the online PMs. 2) Traffic patterns distribution to prevent hotspots in the data center network. Also, the network awareness is achieved via feedback channels of the main network switches.

Network aware server consolidation has been studied also in distributed clouds [43], [44]. From users' point of view, distributed clouds are similar to current cloud providers and provide normal cloud functions and services (i.e. on demand service providing and pay-per-use payment basis). However, in terms of backend implementation, distributed clouds are geographically distributed over a large number of data centers which are connected using a wide area network (WAN) connection [45]. Because of the geographical distribution of the data centers in this case, a user request may have its resource demands from multiple data centers [44]. In data-intensive applications in data centers, a VM which runs an application may run on a PM which is far away from the data center that holds the corresponding data storages. As a result, overall application performance may decrease because of the costly data transfer between the two data centers that hold the VM and data storages [43]. Piao et al. in [43] proposed an algorithm which optimizes the VM placement on a PM such that the data transfer time is minimized. Also, VM Migrations are triggered when data transfer time reaches a threshold determined by SLA. In such case, the VM moves to another PM which will cause the application performance to increase.

When a user requests a service such as mail or social network access, one or more VMs connected together will be assigned to the user. These VMs might be on different PMs and even in different data centers. Obviously, longer distance between VMs leads to a decrease in available bandwidth and an increase in the latency for application tasks to complete, and as a result, the performance will be degraded [44]. Alicherry *et al.* in [44]

proposed a VM to PM mapping method that minimizes the inter-datacenter and intra-datacenter traffic and as a result, the path for transferring packets will be minimized and this leads the application performance to be increased.

C. Considering Cooling Systems

Cooling systems are among energy hungry components of modern data centers which may consume up to 50% of total data center electricity [46]. Hence, reducing the power usage for cooling equipment can significantly reduce the total amount of power usage in a data center. Several studies have been done on cooling aware server consolidation. An important issue in data centers is thermal management which has high effects on cooling energy consumption due to heat recirculation and hotspots [47]. As computation resources processing the received tasks produce heat and their temperature rises, the cooling systems must supply cold air to the server air inlet to cool them down according to server temperature threshold which is the maximum operational temperature that the device can sustain determined by the manufacturer company. The cool air gets into the server from their air inlets and then servers send out the hot air from their outlets. Because of air recirculation, the hot air can turn back into the cold aisle and increases the inlet temperature which can create hotspots [47]. Tang *et al.* in [48] proposed *XInt*, a job scheduling algorithm that minimizes the temperature of inlet air. As a result, the hotspots and heat recirculation impact will be minimized which decreases the cooling equipment energy consumption. Pakbaznia *et al.* in [49] presented a power and thermal management framework, in addition to optimizing servers energy consumption, where the proposed algorithm also minimizes the air conditioning system power consumption by using Dynamic Voltage and Frequency Scaling (DVFS) technique to reduce servers power consumption and also by choosing an optimum temperature for the supplied cold air. Ahmad *et al.* in [50] proposed two algorithms in their work: *PowerTrade* algorithm to tradeoff servers utilization and cooling systems energy consumption which distributes the load over the data center and reduces the power input for cooling systems; Their other proposed algorithm is *SugerGuard* which overprovisions the resources more than required needs to absorb the possible future load fluctuations and request bursts.

D. Considering Performance Impact

The virtualization technology and server consolidation technique introduce a degree of performance interference between VMs which causes an impact on system throughput and overall data center performance [51]. This performance interference can be divided into two major groups which are discussed in the following parts of the paper:

1) *Inter-VM Performance Degradation*

The first factor that decreases the overall application performance is inter-VM performance degradation. Using the consolidation technique, several VMs are packed into PMs. Current virtualization techniques do not guarantee to effectively isolate performance interference between the VMs, and hence, lead to performance interference between them. This will decrease the QoS and may violate the SLA [52]. Also, the resources such as I/O devices, memory capacity, shared cache, shared memory bandwidth, etc., could be affected by this performance interference [53]. However, contention and interference in resources such as memory bandwidth and shared caches can significantly decrease the performance which is measured for several workloads [51], [54], [55]. Also, several studies have been done to isolate the resources such as disk bandwidth [56] and network bandwidth [57]. Thus, performance interference and QoS degradation due to server consolidation are important facts that should be considered when devising server consolidation and resource allocation algorithms.

The proposed algorithms in [54], [58], use the performance profiling method [51], [55], [59] to compute the performance degradation for any possible VM combinations on a PM. Based on that, the VMs that make less performance interference are identified for mapping on a single PM and as a result, in addition to packing the VMs on the lowest possible PMs, the performance degradation will also be minimized. In [60], the authors introduced two approaches named *Performance-Mode* and *Eco-Mode*. For performance prioritized applications, they used *Performance-Mode* which considers a performance bound to ensure every PM meets this minimum bound while minimizing the resource cost and the number of online PMs. Also, they used *Eco-Mode* for jobs that need resource efficiency (e.g. batch processing) which tries to maximize the utilization while minimizing the worst case performance degradation. An important challenge in all these methods, however, is developing performance degradation profiling and prediction methods. Several studies have been done to determine collocated VMs performance impacts [55], [58], [61], [62], [63], [64]; their detailed discussion is beyond the scope of this paper.

2) *Performance Degradation Caused by Software Aging*

In virtualized datacenters, the second factor that decreases the service performance is software aging. Software aging refers to the fact that the system performance faces degradation over the passage of time. The symptoms and effects of this degradation in period of time is data corruption and exhaustion of system resources which leads to performance degradation, software crashes or hanging [65] which can be discovered and fixed in development and test phases of software developing [66]. In long running applications, doing software rejuvenation periodically, decreases the chance of occurring application performance degradation and failures [67]. Since the VM and VMM are both software, they also need various resources such as memory and files. Thus, the software aging in a cloud data center might also occur in either VM or VMM [68]. If the software aging causes a crash or failure in the VMM, all the VMs which are hosted by the VMM will be affected [69]. One of the inexpensive and proactive techniques to absorb this problem is software rejuvenation. This can be done in three ways:

a) *Cold-VM Rejuvenation:* Using this technique, first the administrator resets all the VMs which are on a PM, and then restarts the VMM and the PM and at last starts all the VMs again [69]. This obviously incurs a high service downtime (i.e. the services which are hosted by the VMs), so we need a more efficient way to do the rejuvenation.

b) *Warm-VM Rejuvenation:* Another technique to do the rejuvenation is warm-VM rejuvenation introduced in [69], [70]. In this method, the VMM saves the memory image of all the VMs that it hosts into the hard disk before rebooting and after rebooting it reuses the saved memory images to resume the VMs as they were before. To do this, VMM uses the memory suspend/resume mechanism to suspend the VMs before rebooting, and resumes them after getting online again so as to reduce the service downtime.

c) *Migrate-VM Rejuvenation***:** In [71], authors proposed a technique named Migrate-VM Rejuvenation. In this technique, before the VMM rejuvenation takes place, all the VMs are migrated to another available PM, and then it resets the VMM for rejuvenation.

Authors in [71] compared these three methods. From the steady-state availability prospective, the warm-VM rejuvenation is not always better than cold-VM rejuvenation. Also, they show that migrate-VM rejuvenation is the best technique among these three techniques only if the live migration is fast enough and also data center has enough available resources to host the migrated VMs.

E. Considering Reliability

Reliability in consolidation algorithm could be discussed from two general aspects: Hardware reliability and service reliability. Server consolidation could affect and reduce the reliability and lifetime of the data center devices (e.g. servers). In server consolidation, we try to pack the VMs into fewer number of PMs and turn off the idle PMs. Rapid on-off cycles will reduce the servers' life time. Also, server consolidation increases the servers utilization and as a result the temperature of servers will increase and this can also decrease the servers life time [26]. In addition, hardware failures can lead to service unavailability, SLA violation and performance degradation for the end users [72]. Several studies have been done on reliability aware resource provisioning algorithms. Deng *et al.* in [26] presented a dynamic server consolidation algorithm which considers hardware reliability and lifetime. It uses three parameters U_{SLA} ،U_r and U_e to determine the best VM to PM mapping. U_{SLA} is used to ensure that there are enough resources to support the SLA. U_r holds the value of the impacts of turning servers on and off and temperature changes on reliability and lifetime, and finally U_e shows the amount of power usage reduction for the selected VM to PM mapping. At last, the mapping which has the maximum value of sum of these three parameters is chosen as the optimal mapping. Guenter *et al.* in [73] present *Marlowe*, a service provisioning framework which trades off these three key factors: Cost, performance and reliability. Marlowe predicts the future workload and turns servers on before they are needed. Also, it maximizes energy saving while minimizing the unmet demand and balances the energy savings vs. reliability costs for on-off cycles as well.

Also, service reliability/availability is also an important parameter in data centers which could highly affect the service quality and user experience. In [74], authors considered the service reliability in their algorithm. The proposed algorithm considers power, performance and service reliability aspects altogether. It focuses on two main parts: 1) Guaranteeing average response time for the end users. 2) Using active/active sparing model for servers in which the datacenter uses one of the active servers in a round-robin fashion so as to make the services more stable and reliable.

F. Considering Migration Overheads

A major technology which makes server consolidation even more attractive is VM Live Migration. For live migration, resources (e.g., CPU, memory and network bandwidth) are needed in both source and destination PMs [75]. Network bandwidth usage by VM Live Migration could have negative effects on network efficiency for end users [76] because it uses significant bandwidth for a period of time i.e. 500 Mb/s for 10 seconds for a petty web server as shown in [77]. Depending on the application, the CPU overhead can easily get 30% above the application default CPU demand [78]. Also, if a PM has too many Live Migrations at a moment, this can lead to collision of multiple live migrations [79]. Thus, considering migration needs and overheads in the VM placement algorithm could increase the overall datacenter performance and efficiency. Several studies have been done to address this challenge. Takahashi *et al.* in [79] proposed an algorithm that the number of concurrent live migrations to/from a PM is considered as a limitation in the optimization algorithm. They showed that this will increase the PM throughput and performance. As mentioned before, VM live migration needs some resources in the source and destination PM. Setzer *et al.* in [24] introduced an algorithm in which the resource demands are also considered. They showed that this could prevent the performance degradation and unplanned overloads.

V. CLASSIFICATION BASED ON THE OPTIMIZATION METHOD USED

Apart from the time of applying consolidation and the parameters considered during the decision-making process, the algorithm and method used for smart assignment of VMs to PMs can also be an important factor in final quality of the approach in terms of effectiveness as well as time and resources required to run the algorithm. In this subsection, we review major algorithms and approaches used to solve server consolidation problem.

Generally, there are two types of optimization methods: exact and approximate approaches. Also, approximate methods could be divided in two subgroups: Heuristics and Metaheuristics. Exact optimization methods guarantee finding an optimal solution for the problem, but the time that need to solve the problem will exponentially increase with growth of problem size. So, normally exact methods are used for the problems that belong to class P, or the NP-hard problems with very small

problem size [80]. To overcome these problems, heuristic and metaheuristic optimization methods are used which will be discussed in below subsections.

A. Exact Solutions

In order to determine the optimum VM to PM mapping, first the problem has to be modelled using a mathematical approach, and then effective algorithms must be devised to solve it. Using these exact methods, the optimal solution for a problem could be found. The goal is usually to find a near-optimal solution, not the absolute optimal mapping, since the problem is NP-hard in general [116] and exact methods are so time consuming methods for NP-Hard problems. There are several approaches such as Linear Programming (LP), Dynamic Programming, and Stochastic Programming, which could be used to model the consolidation problem. The VM consolidation problem is mapped into the vector bin-packing classic optimization problem in [81] where the hosts can be considered as bins and VMs are the objects that are going to be packed into the PMs considering limitations some of which we discussed in previous section. There are several widely used ways to solve an NP-Hard problem. Among the most common solutions, heuristic and meta-heuristic methods are general techniques that try to approach the optimal solution by various kinds of intuitions, simple solutions, or inspirations from nature and natural processes of evolution; a number of these techniques are discussed in next subsections. Several exact methods have been used in the literature to formulate the server consolidation problem as seen in Table 1. A large collection of such approaches is described in [82] and [83] for interested readers.

TABLE I. EXACT SOLUTIONS USED IN SERVER CONSOLIDATION ALGORITHMS

Exact Solutions	Used By
Stochastic Programming	Chaisiri *et al.* [84], Ting *et al.* [85]
Linear Programming	T. C. Ferreto *et al.* [18], B. Guenter *et al.* [73], E. Pakbaznia *et al.* [49], [86], A. Pahlavan *et al.* [87], J. Anselmi *et al.* [88]
Non-Linear Programming	A. Sansottera [74], J. Anselmi *et al.* [88]
Dynamic Programming	H. Goudarzi [89], [90], J.J. Wu *et al.* [91]
Constraint Programming	F. Hermenier [92], K. Dhyani *et al.* [93]
Quadratic Programming	O. Biran [94]
Game Theory	F. Teng *et al.* [95], S. U. Khan *et al.* [96]

B. Heuristics

Server consolidation is a multidimensional bin packing problem which is an NP-Hard optimization problem in core [81]. Heuristics are problem-dependent methods that although do not guarantee finding the optimal solution, but they try to find a near-optimal solution in a reasonably and practically short time. Heuristic optimization methods showed a good performance for solving the NP-Hard problems. So, because of high complexity of the server consolidation problem, and also data center real-time operation, heuristics are good methods to use and solve the server consolidation optimization problem. There are several heuristics to help find a solution for the server consolidation problem. One of the most popular ones that is basically a locally optimal algorithm, is a greedy algorithm named First Fit Decreasing (FFD) [97], [94]. This algorithm sorts the VMs in decreasing order of resource demands and then maps the VMs from the top of the list onto the first PM which has enough capacity in terms of resources. A major limitation of this heuristic is that the problem must be one-dimensional and also the PMs must have the same resource capacity [81]. The advantage of this method is simplicity and speed, but it cannot guarantee to find the most appropriate VM to PM mapping. Another common heuristic is Best Fit Decreasing (BFD) [9]. BFD first sorts the VMs based on their resource demands in decreasing order, and then allocates the VMs to the PM with resources closest to the VM requirements. There are many comprehensive books and papers such as [81], [98], [99] that describe the theory and several heuristics on server consolidation and bin-packing problems for further reading on this topic. These famous heuristics that are used in solving combinatorial problems (e.g. bin packing) are listed and available in table 2. However, because of increasing complexity of the consolidation problem, authors mostly prefer to propose their own heuristic algorithms to solve the problem.

TABLE II. FAMOUS HEURISTICS USED TO SOLVE THE BIN PACKING PROBLEM AND SERVER CONSOLIDATION PROBLEMS

Heuristics	Used By
First Fit Decreasing (FFD)	N. Bobroff *et al.* [97], A. Verma *et al.*[30]
Best Fit Decreasing (BFD)	J. Xue *et al.* [100], T. C. Ferreto *et al.* [18]
Next Fit	K. Mills *et al.* [101], M. Wang *et al. [39]*
Random Fit	K. Mills *et al.* [101]
Least Full First	Y. Ajiro *et al.* [98]
Most Full First	Lee *et al.* [81]
Dot Product	M. Mishra *et al.* [102]
Minimizing Angle	M. Mishra *et al.* [102]

C. Meta-Heuristics

Another approximate optimization method which is widely used to solve the optimization problems is meta-heuristics. Meta-heuristics as opposed to heuristics are problem-independent techniques. Metaheuristics are strategies that effectively guides the space search process in order to find (near-) optimal solutions and usually takes more time than quick heuristics to find the solution [103]. There are various metaheuritics that have been used to solve the server consolidation problem. Table 4 provides a list of some of the mostly used metaheuristics along with the papers that used them to solve the optimization problem. Obviously these algorithms show different benefits and performances based on the problem and the test bed. Metaheuristic algorithms [104] and their usage in resource management and scheduling problems are comprehensively explained in [105] for interested readers.

TABLE III. COMPARISON OF SERVER CONSOLIDATION APPROACHES

Paper	Objective(s)	Time of Decision Making	Hardware Considered	Performance	Reliability		Considered Cooling System	Considered Migration Overhead	Considered Network Traffic	Evaluation
					Service Reliability	Hardware Reliability				
[28]	Minimize resource waste Minimize SLA violations	Dynamic with Load Prediction	CPU	No	No	No	No	No	No	Practical
[37]	Maximize the network scalability	Dynamic	CPU, Memory	No	No	No	No	No	Yes	Simulation
[106]	Minimize the total amount of resources cost Meet end-to-end performance requirements	Dynamic	CPU, Memory	Yes	No	No	No	No	No	Simulation
[42]	Minimize number of active hosts Avoid hotspots in the data center network	Dynamic	CPU	No	No	No	No	No	Yes	Simulation
[44]	Minimize the maximum distance (or latency) in distributed clouds	Dynamic	CPU	No	No	No	No	No	Yes	Simulation
[48]	Minimize the inlet temperatures (leads to lower energy consumption)	Dynamic	CPU	No	No	No	Yes	No	No	Simulation
[49]	Minimize the total data center power cost (Cooling and server)	Dynamic	CPU, Memory	No	No	No	Yes	No	No	Simulation
[50]	Trades-off idle power and cooling power for each other, thereby reducing the total power	Dynamic	CPU	Yes	No	No	Yes	No	No	Simulation
[94]	Maximize server utilization Minimize migration overhead	Dynamic	CPU, Memory, Bandwidth	Yes	No	No	No	Yes	No	Simulation
[60]	Minimize resource cost (e.g. energy) Minimize worst case performance degradation	Dynamic with Load Prediction	CPU, Memory, Disk	Yes	No	No	No	No	No	Simulation
[26]	Maximize hardware lifetime Minimize energy consumption	Dynamic with Load Prediction	CPU, Memory, Disk, Network	Yes	No	Yes	No	Yes	No	Simulation
[73]	Maximize energy savings Minimizing unmet demand Minimize reliability costs of server components	Dynamic with Load Prediction	CPU	Yes	No	Yes	No	No	No	Simulation
[107]	Minimize total energy cost while meeting the performance objective	Dynamic	CPU, Memory	Yes	No	No	No	No	No	Simulation
[74]	Maximize cost savings Minimize SLA violations Maximize service reliability	Dynamic	CPU	Yes	Yes	No	No	No	No	Practical
[79]	Minimize power consumption Minimize SLA violations	Dynamic	CPU	Yes	No	No	No	Yes	No	Simulation
[24]	Minimize number of active hosts	Static	CPU	Yes	No	No	No	Yes	No	Simulation
[36]	Maximize resource utilization Minimize SLA violations	Dynamic with Load Prediction	CPU, Memory, Bandwidth	Yes	No	No	No	No	No	Simulation
[35]	Maximize resource utilization	Dynamic	CPU, Memory, Disk	No	No	No	No	No	No	Simulation

TABLE IV. METAHEURISTICS USED TO SOLVE THE SERVER CONSOLIDATION PROBLEMS

Metaheuristic Methods	Used By
Genetic Algorithm (GA) [108]	J.J. Prevost *et al.* [23], H. Hlavacs *et al.[109]*, Ligang He *et al.* [110], A. C. Adamuthe [111], Mehdi *et al.* [112]
Grouping Genetic Algorithm (GGA) [113]	W. Deng *et al.[26]*, S. Agrawal *et al.* [114], D. Wilcox *et al.* [115]
Ant Colony (ACO) [116]	X. F. Liu *et al.*, Y Gao *et al.* [117], A. Ashraf *et al.* [118], M. H. Ferdaus *et al.* [119], Sarma *et al.* [120], G. Xu *et al. [121]*
Simulated Annealing (SA) [122]	Y. Wu *et al.*[123], P. Zhang *et al.* [124]
Particle Swarm Optimization (PSO) [125]	A. C. Adamuthe [111], C. C. T. Mark *et al.* [126]
Tabu Search [127]	T. Ferreto *et al.* [128]
Hybrid Optimization [129]	C. C. T. Mark *et al.* [126], J. Dong *et al.* [130], B. B. J. Suseela *et al.* [131]

VI. CONCLUSION AND OPEN CHALLENGES

This survey discussed server consolidation techniques for reducing datacenter energy consumption as an important challenge for sustainable development of internet-scale IT systems and services in both industry and academia. Our brief review of cloud data centers and optimization opportunities provided by virtualization technology constructed the background required to understand the rest of the paper contents and their significance. We then presented a system model and reviewed various approaches to server consolidation presented in the literature and classified them from five points of view: time of applying the technique, constraints and requirements considered during optimization process, and algorithmic method used to find near-optimal solution of the optimization problem, their objective functions and evaluation methods. .

Generally, there are two phases in developing server consolidation approaches. Phase 1: Problem definition (i.e. objective function(s) and constraints). Phase 2: Solving the optimization problem using different techniques. There are many challenges and future works to explore in both phases. We provide our view of open challenges in the following two subsections.

A. Problem Definition Phase

While early-bird techniques that focus mainly on a few constraints, such as merely CPU utilization, seem saturated in the literature, there is ample space for holistic approaches that simultaneously consider all or multiple resources including CPU, memory, disk, and network bandwidth. Developing consolidation approaches considering multiple system resources for optimization could lead to more efficient and applicable approaches.

Data storage and networking equipment are two important parts that have received less attention in previous works. Some data centers use centralized (e.g. Storage Area Network storage systems) [ref] and some others benefit from distributed storage in form of local disks on servers [ref]. These two approaches reflect different behaviors in terms of energy consumption when it comes to VM consolidation and VM live migration; this happens since the amount of data, including OS image as well as applications data, to be transferred over the data center network differ. This is another area that has been overlooked up to now although storage-only or network-only awareness has been covered before.

A downside of VM consolidation largely overlooked up to now is that once a PM fails, all VMs running on it will fail. Thus, the interaction between dependability of the service, the failure recovery techniques it uses, and total energy consumption of the scheme is another interesting tradeoff to explore. One can add thermal and heating effects of servers, and even more importantly of routers, to this to make it an even more challenging problem to attack.

Performance behavior of applications, especially their interference on one another when consolidated as VMs on a PM is deeper than current works have explored. While prior works basically consider the interferences to be static and known a priori, run-time behavior of applications especially in modern data centers where several users' VMs with different applications (e.g. scientific, social networking, enterprise applications) could be allocated on a single PM, are inherently more complex and changing over time. Automatic determination of compatible applications and efficient allocation and parameter tuning of them on a single node is an important challenge to explore.

Although all consolidation objectives are important, but some of them are in contrast to each other. For example, increasing resource utilization is in contrast to minimizing heat and cooling efficiency. Therefore, as shown in table 3, different consolidation strategies need to be combined to satisfy multiple objectives at the same time which increases challenge and complexity: performance, power, total cost of service provisioning, availability of the service, reliability and life time of components, dynamic nature of usage of various resources such as CPU, memory, and network, and their potentially periodic repetition on a daily basis due to the nature of internet-scale services, are only a few samples of important objectives and/or constraints that need to be considered.

Applying various predictions is another avenue to explore. Predictive methods to forecast future needs of the VM so as to resize or migrate them in time, or to prevent repetitive on-off cycles of the PM still have potential for further work. Prediction methods are also required to estimate performance impact of collocating a number of given VMs.

B. Problem Solving Phase

Due to the ever increasing expansion of internet-scale online services such as social networks, data centers are becoming larger in size and quantity which puts more pressure on consolidation techniques for response time and scalability. Therefore, there is an increasing need for approaches that are decentralized, hierarchical, and fast. Here, the research challenges are: How to effectively combine different optimization techniques? Can hierarchal approaches and hybrid metaheuristics help on this? What is the optimal time period for running the algorithm? How parallel algorithm approaches can help to speed up the algorithm run time?

In terms of computing system design we found very few works to cite in this paper whereas one can think of many improvement opportunities in this scope. Memory architecture of servers to support faster hibernation and rejuvenation of VMs, less inter-VM performance conflict when mapped to the same PM, and cross-layer hardware-software collaborative techniques for more efficient and near-optimal resource sharing among VMs and considering its effect on consolidation choices are among them only to name a few.

REFERENCES

[1] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, *et al.*, "Above the clouds: A Berkeley view of cloud computing," *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS,* vol. 28, p. 13, 2009.

[2] *Amazon EC2*. Available: aws.amazon.com/ec2/

[3] *Microsoft Windows Azure*. Available: https://azure.microsoft.com/

[4] *Google Cloud*. Available: https://cloud.google.com/

[5] *IBM Cloud Service*. Available: www-935.ibm.com/services/us/en/it-services/cloud-services/

[6] R. Zhu, Z. Sun, and J. Hu, "Special section: Green computing," *Future Generation Computer Systems,* vol. 28, pp. 368-370, 2012.

[7] *World Energy Outlook 2013 Fact Sheet*. Available: http://www.iea.org/media/files/WEO2013_factsheets.pdf

[8] W. Wang, H .Chen, and X. Chen, "An availability-aware virtual machine placement approach for dynamic scaling of cloud applications," in *Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on*, 2012, pp. 509-516.

[9] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generation Computer Systems,* vol. 28, pp. 755-768, 2012.

[10] R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges," *arXiv preprint arXiv:1006.0308,* 2010.

[11] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *IEEE computer,* vol. 40, pp. 33-37, 2007.

[12] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, "Optimal power allocation in server farms," in *ACM SIGMETRICS Performance Evaluation Review*, 2009, pp. 157-168.

[13] M. Cardosa, M. R. Korupolu, and A. Singh, "Shares and utilities based power consolidation in virtualized server environments," in *Integrated Network Management, 2009. IM'09. IFIP/IEEE International Symposium on*, 2009, pp. 327-334.

[14] A. Wolke, M. Bichler, and T. Setzer, "Planning vs. dynamic control: Resource allocation in corporate clouds," *IEEE Transactions of Cloud Computing*.

[15] *Citrix Xen*. Available: http://www.citrix.com/products/xenserver/

[16] *VMware vCenter Server*. Available: http://www.vmware.com/products/vcenter-server/

[17] *Server Consolidation and Virtualization Analysis by CiRBA*. Available: http://www.cirba.com/

[18] T. C. Ferreto, M. A. Netto, R. N. Calheiros, and C. A. De Rose, "Server consolidation with migration control for virtualized data centers," *Future Generation Computer Systems,* vol. 27, pp. 1027-1034, 2011.

[19] K. Halder, U. Bellur, and P. Kulkarni, "Risk aware provisioning and resource aggregation based consolidation of virtual machines," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, 2012, pp. 598-605.

[20] B. Speitkamp and M. Bichler, "A mathematical programming approach for server consolidation problems in virtualized data centers," *Services Computing, IEEE Transactions on,* vol. 3, pp. 266-278.2010 ,

[21] A. Verma, G. Dasgupta, T. K. Nayak, P. De, and R. Kothari, "Server workload analysis for power minimization using consolidation," in *Proceedings of the 2009 conference on USENIX Annual technical conference*, 2009, pp. 28-28.

[22] G. Lovász, F .Niedermeier, and H. de Meer, "Performance tradeoffs of energy-aware virtual machine consolidation," *Cluster Computing,* vol. 16, pp. 481-496, 2013.

[23] J. J. Prevost, K. Nagothu, B. Kelley, and M. Jamshidi, "Optimal update frequency model for physical machine state change and virtual machine placement in the cloud," in *System of Systems Engineering (SoSE), 2013 8th International Conference on*, 2013, pp. 159-164.

[24] T. Setzer and A. Wolke, "Virtual machine re-assignment considering migration overhead," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*, 2012, pp. 631-634.

[25] V. Ebrahimirad, M. Goudarzi, and A. Rajabi, "Energy-Aware Scheduling for Precedence-Constrained Parallel Virtual Machines in Virtualized Data Centers," *Journal of Grid Computing,* vol. 13, pp. 233-253, 2015.

[26] W. Deng, F. Liu, H. Jin, X. Liao, and H. Liu, "Reliability-aware server consolidation for balancing energy-lifetime tradeoff in virtualized cloud datacenters," *International Journal of Communication Systems,* vol. 27, pp. 623-642, 2014.

[27] W. Xu, X. Zhu, S. Singhal, and Z. Wang, "Predictive control for dynamic resource allocation in enterprise data centers," in *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*, 2006, pp. 115-126.

[28] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*, 2010, pp. 9-16.

[29] C. Mastroianni, M. Meo, and G. Papuzzo, "Self-economy in cloud data centers: Statistical assignment and migration of virtual machines," in *Euro-Par 2011 Parallel Processing*, ed: Springer, 2011, pp. 407-418.

[30] A. Verma, P. Ahuja, and A. Neogi, "pMapper: power and migration cost aware application placement in virtualized systems," in *Middleware 2008*, ed: Springer, 2008, pp. 243-264.

[31] Y. Song, H. Wang, Y. Li, B. Feng, and Y. Sun, "Multi-tiered on-demand resource scheduling for VM-based data center," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2009, pp. 148-155.

[32] D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper, "Resource pool management: Reactive versus proactive or let's be friends," *Computer Networks,* vol. 53, pp. 2905-2922, 2009.

[33] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, 2010, pp. 826-831.

[34] A. Fox, A. Turner, and H. S. Kim, "Resource contention-aware Virtual Machine management for enterprise applications," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, 2012, pp. 1641-1646.

[35] Y. C. Lee and A. Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems," *The Journal of Supercomputing,* vol. 60, pp. 268-280, 2012.

[36] S. K. Garg, A. N. Toosi, S. K. Gopalaiyengar, and R. Buyya, "SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter," *Journal of Network and Computer Applications,* vol. 45, pp. 108-120, 2014.

[37] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *INFOCOM, 2010 Proceedings IEEE*, 2010, pp. 1-9.

[38] T. Benson, A. Anand ,A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *ACM SIGCOMM Computer Communication Review,* vol. 40, pp. 92-99, 2010.

[39] M. Wang, X. Meng, and L. Zhang, "Consolidating virtual machines with dynamic bandwidth demand in data centers," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 71-75.

[40] O. Biran, A. Corradi, M. Fanelli, L. Foschini, A. Nus, D. Raz, *et al.*, "A stable network-aware vm placement for cloud systems," in *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, 2012, pp. 498-506.

[41] B. Furht and A. Escalante, *Handbook of data intensive computing*: Springer, 2011.

[42] D. Kliazovich, P. Bouvry, and S. U. Khan, "DENS: data center energy-efficient network-aware scheduling," *Cluster computing,* vol. 16, pp. 65-75, 2013.

[43] J. T. Piao and J. Yan, "A network-aware virtual machine placement and migration approach in cloud computing," in *Grid and Cooperative Computing (GCC), 2010 9th International Conference on* ,2010 ,pp. 87-92.

[44] M. Alicherry and T. Lakshman, "Network aware resource allocation in distributed clouds," in *INFOCOM, 2012 Proceedings IEEE*, 2012, pp. 963-971.
[45] P. T. Endo, A. V. de Almeida Palhares, N. N. Pereira, G. E. Goncalves, D. Sadok, J. Kelner*, et al.*, "Resource allocation for distributed cloud: concepts and research challenges," *Network, IEEE,* vol. 25, pp. 42-46, 2011.
[46] R. Sawyer, "Calculating total power requirements for data centers," *American Power Conversion, Tech. Rep,* vol. 70 ,pp. 80-90, 2004.
[47] C. Bash and G. Forman, "Cool Job Allocation: Measuring the Power Savings of Placing Jobs at Cooling-Efficient Locations in the Data Center," in *USENIX Annual Technical Conference*, 2007, p. 140.
[48] Q. Tang, S. K. Gupta, and G. Varsamopoulos, "Thermal-aware task scheduling for data centers through minimizing heat recirculation," in *Cluster Computing, 2007 IEEE International Conference on*, 2007, pp. 129-138.
[49] E. Pakbaznia and M. Pedram, "Minimizing data center cooling and server power costs," in *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, 2009, pp. 145-150.
[50] F. Ahmad and T. Vijaykumar, "Joint optimization of idle and cooling power in data centers while maintaining response time ",in *ACM Sigplan Notices*, 2010, pp. 243-256.
[51] Y. Koh, R. C. Knauerhase, P. Brett, M. Bowman, Z. Wen, and C. Pu, "An Analysis of Performance Interference Effects in Virtual Environments," in *ISPASS*, 2007, pp. 200-209.
[52] R. B. Nathuji and A. Ghaffarkhah, "Managing performance interference effects on cloud computing servers," ed: Google Patents, 2013.
[53] O. Tickoo, R. Iyer, R. Illikkal, and D. Newell, "Modeling virtual machine performance: challenges and approaches," *ACM SIGMETRICS Performance Evaluation Review,* vol. 37, pp. 55-60, 2010.
[54] Y. Jiang, X. Shen, J. Chen, and R. Tripathi, "Analysis and approximation of optimal co-scheduling on chip multiprocessors," in *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*, 2008, pp. 220-229.
[55] J. Mars, L. Tang, R. Hundt, K. Skadron, and M. L. Soffa, "Bubble-up: Increasing utilization in modern warehouse scale computers via sensible co-locations," in *Proceedings of the 44th annual IEEE/ACM International Symposium on Microarchitecture*, 2011, pp. 248-259.
[56] M. K. Qureshi and Y. N. Patt, "Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches," in *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, 2006, pp. 423-432.
[57] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat, "Enforcing performance isolation across virtual machines in Xen," in *Middleware 2006*, ed: Springer, 2006, pp. 342-362.
[58] Y. Jiang, K. Tian, and X. Shen, "Combining locality analysis with online proactive job co-scheduling in chip multiprocessors," in *High performance embedded architectures and compilers*, ed: Springer, 2010, pp. 201-215.
[59] J. Du, N. Sehrawat, and W. Zwaenepoel, "Performance profiling of virtual machines," *ACM SIGPLAN Notices,* vol. 46, pp. 3-14, 2011.
[60] A. Roytman, A. Kansal, S. Govindan, J. Liu, and S. Nath, "PACMan: Performance Aware Virtual Machine Consolidation," in *ICAC*, 2013, pp. 83-94.
[61] C. Delimitrou and C .Kozyrakis, "Paragon: QoS-aware scheduling for heterogeneous datacenters," *ACM SIGARCH Computer Architecture News,* vol. 41, pp. 77-88, 2013.
[62] J. Mars, L. Tang, and M. L. Soffa, "Directly characterizing cross core interference through contention synthesis," in *Proceedings of the 6th International Conference on High Performance and Embedded Architectures and Compilers*, 2011, pp. 167-176.
[63] J. Zhao, H. Cui, J. Xue, X. Feng, Y. Yan, and W. Yang, "An empirical model for predicting cross-core performance interference on multicore processors," in *Proceedings of the 22nd international conference on Parallel architectures and compilation techniques*, 2013, pp. 201-212.
[64] L. Y. Chen, G. Serazzi, D. Ansaloni, E. Smirni, and W. Binder, "What to expect when you are consolidating: effective prediction models of application performance on multicores," *Cluster computing,* vol. 17, pp. 19-37, 2014.
[65] V. Castelli, R. E. Harper, P. Heidelberger, S. W. Hunter, K. S. Trivedi, K. Vaidyanathan*, et al.*, "Proactive management of software aging," *IBM Journal of Research and Development,* vol. 45, pp. 311-332, 2001.
[66] K. Trivedi, G. Ciardo, B. Dasarathy, M. Grottke, R. Matias, A. Rindos*, et al.*, "Achieving and assuring high availability," in *Service Availability*, ed: Springer, 2008, pp. 20-25.
[67] Y. Huang, C. Kintala, N. Kolettis, and N. D. Fulton, "Software rejuvenation: Analysis, module and applications," in *Fault-Tolerant Computing, 1995. FTCS-25. Digest of Papers., Twenty-Fifth International Symposium on*, 1995, pp.390-381 .
[68] F. Machida, D. S. Kim, J. S. Park, and K. S. Trivedi, "Toward optimal virtual machine placement and rejuvenation scheduling in a virtualized data center," in *Software Reliability Engineering Workshops, 2008. ISSRE Wksp 2008. IEEE International Conference on*, 2008, pp. 1-3.
[69] K. Kourai and S. Chiba, "Fast software rejuvenation of virtual machine monitors," *Dependable and Secure Computing, IEEE Transactions on,* vol. 8, pp. 839-851, 2011.
[70] K. Kourai and S. Chiba, "A fast rejuvenation technique for server consolidation with virtual machines," in *Dependable Systems and Networks, 2007. DSN'07. 37th Annual IEEE/IFIP International Conference on*, 2007, pp. 245-255.
[71] F. Machida, D. S. Kim, and K. S. Trivedi, "Modeling and analysis of software rejuvenation in a server virtualized system," in *Software Aging and Rejuvenation (WoSAR), 2010 IEEE Second International Workshop on*, 2010, pp. 1-6.
[72] D. Oppenheimer, A. Ganapathi, and D. A. Patterson, "Why do Internet services fail, and what can be done about it?," in *USENIX Symposium on Internet Technologies and Systems*, 2003.
[73] B. Guenter, N. Jain, and C. Williams, "Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning," in *INFOCOM, 2011 Proceedings IEEE* ,2011 ,pp. 1332-1340.
[74] A. Sansottera, D. Zoni, P. Cremonesi, and W. Fornaciari, "Consolidation of multi-tier workloads with performance and reliability constraints," presented at the High Performance Computing and Simulation (HPCS), 2012 International Conference on, Madrid, Spain, 2012.
[75] J. Hall, J. Hartline, A. R. Karlin, J. Saia, and J. Wilkes, "On algorithms for efficient data migration," in *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, 2001, pp. 620-629.
[76] A .Stage and T. Setzer, "Network-aware migration control and scheduling of differentiated virtual machine workloads," in *Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, 2009, pp. 9-14.
[77] C. Clark, K. Fraser, S .Hand, J. G. Hansen, E. Jul, C. Limpach*, et al.*, "Live migration of virtual machines," in *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*, 2005, pp. 273-286.
[78] S. Akoush, R. Sohan, A. Rice, A. W. Moore, and A. Hopper, "Predicting the performance of virtual machine migration," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2010 IEEE International Symposium on*, 2010, pp. 37-46.
[79] S. Takahashi, A. Takefusa, M .Shigeno, H. Nakada, T. Kudoh, and A. Yoshise, "Virtual machine packing algorithms for lower power consumption," in *High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:*, 2012, pp. 1517-1518.
[80] F. Rothlauf, *Design of modern heuristics: principles and application*: Springer Science & Business Media, 2011.
[81] S. Lee, R. Panigrahy, V. Prabhakaran, V. Ramasubramanian, K. Talwar, L. Uyeda*, et al.*, "Validating heuristics for virtual machines consolidation," *Microsoft Research, MSR-TR-2011-9,* 2011.
[82] C. A. Floudas and P. M. Pardalos, *Encyclopedia of optimization* vol. 1: Springer, 2008.
[83] S. S. Rao and S. Rao, *Engineering optimization: theory and practice*: John Wiley & Sons, 2009.
[84] S. Chaisiri, B.-S. Lee, and D. Niyato, "Optimization of resource provisioning cost in cloud computing," *Services Computing, IEEE Transactions on,* vol. 5, pp. 164-177, 2012.
[85] T. He, S. Chen, H. Kim, L. Tong, and K.-W. Lee, "Scheduling parallel tasks onto opportunistically available cloud resources," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, 2012, pp. 180-187.
[86] E. Pakbaznia, M. Ghasemazar, and M. Pedram, "Temperature-aware dynamic resource provisioning in a power-optimized datacenter," in *Proceedings of the Conference on Design, Automation and Test in Europe*, 2010, pp. 124-129.
[87] A. Pahlavan, M. Momtazpour, and M. Goudarzi, "Data center power reduction by heuristic variation-aware server placement and chassis

consolidation," in *Computer Architecture and Digital Systems (CADS), 2012 16th CSI International Symposium on*, 2012, pp. 150-155.
[88] J. Anselmi, E. Amaldi, and P. Cremonesi, "Service consolidation with end-to-end response time constraints," in *Software Engineering and Advanced Applications, 2008 .SEAA'08. 34th Euromicro Conference*, 2008, pp. 345-352.
[89] H. Goudarzi and M. Pedram, "Energy-efficient virtual machine replication and placement in a cloud computing system," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, 2012 ,pp. 750-757.
[90] H. Goudarzi, M. Ghasemazar, and M. Pedram, "Sla-based optimization of power and migration cost in cloud computing," in *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*, 2012, pp. 172-179.
[91] J-.J. Wu, P. Liu, and J.-S. Yang, "Workload characteristics-aware virtual machine consolidation algorithms," in *Proceedings of the 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, 2012, pp. 42-49.
[92] F. Hermenier, X. Lorca, J.-M. Menaud, G. Muller, and J. Lawall, "Entropy: a consolidation manager for clusters," in *Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments*, 2009, pp. 41-50.
[93] K. Dhyani, S. Gualandi, and P. Cremonesi, "A constraint programming approach for the service consolidation problem," in *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, ed: Springer, 2010, pp. 97-101.
[94] T. Wood, P. J. Shenoy, A. Venkataramani, and M. S. Yousif, "Black-box and Gray-box Strategies for Virtual Machine Migration," in *NSDI*, 2007, pp. 17-17.
[95] F. Teng and F. Magoulès, "A new game theoretical resource allocation algorithm for cloud computing," in *Advances in Grid and Pervasive Computing*, ed: Springer, 2010, pp. 321-330.
[96] S. U. Khan and I. Ahmad, "A cooperative game theoretical technique for joint optimization of energy consumption and response time in computational grids," *Parallel and Distributed Systems ,IEEE Transactions on,* vol. 20, pp. 346-360, 2009.
[97] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on*, 2007 , pp. 119-128.
[98] Y. Ajiro and A. Tanaka, "Improving packing algorithms for server consolidation," in *Int. CMG Conference*, 2007, pp. 399-406.
[99] D. S. Hochba, "Approximation algorithms for NP-hard problems," *ACM SIGACT News,* vol. 28, pp. 40-52, 1997.
[100] J. Xu and J. A. Fortes, "Multi-objective virtual machine placement in virtualized data center environments," in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, 2010, pp. 179-188.
[101] K. Mills, J. Filliben, and C. Dabrowski, "Comparing vm-placement algorithms for on-demand clouds," in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, 2011, pp. 91-98.
[102] M. Mishra and A. Sahoo, "On theory of vm placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, 2011, pp. 275-282.
[103] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," *ACM Computing Surveys (CSUR),* vol. 35, pp. 268-308, 2003.
[104] I. H. Osman and G. Laporte, "Metaheuristics: A bibliography," *Annals of Operations Research,* vol ,63 .pp. 511-623, 1996.
[105] F. Xhafa and A. Abraham, "Meta-heuristics for grid scheduling problems," in *Metaheuristics for Scheduling in Distributed Computing Environments*, ed: Springer, 2008, pp. 1-37.
[106] P. Xiong, Z. Wang, S. Malkowski, Q. Wang, D .Jayasinghe, and C. Pu, "Economical and robust provisioning of n-tier cloud workloads: A multi-level control approach," in *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, 2011, pp. 571-580.
[107] Q. Zhang, M. F. Zhani, S. Zhang, Q. Zhu, R. Boutaba, and J. L. Hellerstein, "Dynamic energy-aware capacity provisioning for cloud computing environments," in *Proceedings of the 9th international conference on Autonomic computing*, 2012, pp. 145-154.
[108] D. E. Goldberg, *Genetic algorithms*: Pearson Education India, 2006.
[109] H. Hlavacs and T. Treutner, "Genetic algorithms for energy efficient virtualized data centers," in *Network and service management (cnsm), 2012 8th international conference and 2012 workshop on systems virtualiztion management (svm)*, 2012, pp. 422-429.
[110] L. He, D. Zou, Z. Zhang, H. Jin, K. Yang, and S. A. Jarvis, "Optimizing resource consumptions in clouds," in *Grid Computing (GRID), 2011 12th IEEE/ACM International Conference on*, 2011, pp. 42-49.
[111] A. C .Adamuthe, V. Bhise, and G. Thampi, "Solving resource provisioning in cloud using GAs and PSO," in *Engineering (NUiCONE), 2013 Nirma University International Conference on*, 2013, pp. 1-5.
[112] N. A. Mehdi, A. Mamat, H. Ibrahim, and S. K. Subramaniam, "Impatient task mapping in elastic cloud using genetic algorithm," *Journal of Computer Science,* vol. 7, p. 877, 2011.
[113] E. Falkenauer, *Genetic algorithms and grouping problems*: John Wiley & Sons, Inc., 1998.
[114] S. Agrawal, S. K. Bose, and S. Sundarrajan, "Grouping genetic algorithm for solving the serverconsolidation problem with conflicts," in *Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation*, 2009, pp. 1-8.
[115] D. Wilcox, A. McNabb, and K. Seppi, "Solving virtual machine packing with a reordering grouping genetic algorithm," in *Evolutionary Computation (CEC), 2011 IEEE Congress on*, 2011, pp. 362-369.
[116] M. Dorigo and G. Di Caro, "Ant colony optimization: a new meta-heuristic," in *Proceedings of the 1999 congress on evolutionary computation*, 1999, pp. 1470-1477.
[117] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journal of Computer and System Sciences,* vol. 79, pp. 123.2013 ,1242-0
[118] A. Ashraf and I. Porres, "Using Ant Colony System to Consolidate Multiple Web Applications in a Cloud Environment," in *Parallel, Distributed and Network-Based Processing (PDP), 2014 22nd Euromicro International Conference on*, 2014, pp.489-482 .
[119] M. H. Ferdaus, M. Murshed, R. N. Calheiros, and R. Buyya, "Virtual Machine Consolidation in Cloud Data Centers Using ACO Metaheuristic," in *Euro-Par 2014 Parallel Processing*, ed: Springer, 2014, pp. 306-317.
[120] V. A. K. Sarma, R. Rajendra, P. Dheepan, and K. S. Kumar, "An Optimal Ant Colony Algorithm for Efficient VM Placement," *Indian Journal of Science and Technology,* vol. 8, pp. 156-159, 2015.
[121] G. Xu, Y. Dong, and X. Fu, "VMs Placement Strategy based on Distributed Parallel Ant Colony Optimization Algorithm," *Appl. Math,* vol. 9, pp. 873-881, 2015.
[122] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simmulated annealing," *science,* vol. 220, pp. 671-680, 1983.
[123] Y. Wu, M. Tang, and W. Fraser, "A simulated annealing algorithm for energy efficient virtual machine placement," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, 2012, pp. 1245-1250.
[124] P. Zhang, H. Wang, J. Dong, Y. Li, and S. Cheng, "SmartShuffle: Managing Online Virtual Machine Shuffle in Virtualized Data Centers," in *Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference on*, 2013, pp. 113-118.
[125] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of Machine Learning*, ed: Springer, 2010, pp. 760-766.
[126] C. C. T. Mark, D. Niyato, and T. Chen-Khong, "Evolutionary optimal virtual machine placement and demand forecaster for cloud computing," in *Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on*, 2011, pp. 348-355.
[127] F. Glover, "Future paths for integer programming and links to artificial intelligence," *Computers & Operations Research,* vol. 13, pp. 533-549, 1986.
[128] T. Ferreto, C. A. De Rose, and H.-U. Heiss, "Maximum migration time guarantees in dynamic server consolidation for virtualized data centers," in *Euro-Par 2011 Parallel Processing*, ed: Springer, 2011, pp. 443-454.
[129] C. Blum, J. Puchinger, G. R. Raidl, and A. Roli, "Hybrid metaheuristics in combinatorial optimization: A survey," *Applied Soft Computing,* vol. 11, pp. 4135-4151, 2011.
[130] J. Dong, X. Jin, H. Wang, Y. Li, P. Zhang, and S. Cheng, "Energy-saving virtual machine placement in cloud data centers," in *Cluster, Cloud and Grid Computing (CCGrid), 2013*13 *th IEEE/ACM International Symposium on*, 2013, pp. 618-624.
[131] B. B. J. Suseela and V. Jeyakrishnan, "A multi-objective hybrid ACO-PSO optimization algorithm for virtual machine placement in cloud computing".